

NOMINAL ASSOCIATION VECTOR AND MATRIX

BY WENXUE HUANG, YONG SHI AND XIAOGANG WANG

Shantou University, Chinese Academy of Science and York University

When response variables are nominal and populations are cross-classified with respect to multiple polytomies, questions often arise about the degree of association of the responses with explanatory variables. When populations are known, we introduce a nominal association vector and matrix to evaluate the dependence of a response variable with an explanatory variable. These measures provide detailed evaluations of nominal associations at both local and global levels. We also define a general class of global association measures which embraces the well known association measure by Goodman-Kruskal (1954). The proposed association matrix also gives rise to the expected generalized confusion matrix in classification. The hierarchy of equivalence relations defined by the association vector and matrix are also shown.

1. Introduction. Many studies in biology, psychology, sociology and text mining deal with nominal dependent response variable with categorical explanatory variables. When a parametric models, such as a logistic or log-linear model, are employed, a standard statistical analysis can be performed to determine the significance of the explanatory variable. Agresti ([1]) and Fienberg ([3]) provide excellent accounts of parametric methods for analyzing categorical data.

AMS 2000 subject classifications: Primary 62H20

Keywords and phrases: Association matrix, association measures, association vector, categorical data, equivalence relation, generalized confusion matrix, the Goodman-Kruskal τ .

When the number of covariates is large, a direct employment of a predictive model could encounter serious computational difficulties. To reduce the dimensionality, one must first select a collection of highly relevant covariates with a much smaller dimensionality. Therefore an accurate evaluation of any existing association among categorical variables becomes crucial for analyzing high dimensional categorical data. A measure of association for categorical variables is referred to as nominal if any possible scale or order of the categories of variables is not of interest. There are several nominal association measures available in the literature. Goodman and Krusal ([7]) argued that many such measures of association for nominal data stem from the standard chi-square statistic upon which a test of independence is usually based. They also argued that the class of measures based on chi-square lacks of interpretability. They considered alternative measures based on proportional predictions. The method of proportional prediction has been widely used in clinic diagnosis, inventory management, risk management and social studies. Goodman ([5] and [6]) provided a thorough review of analysis of cross-classified data and also presented a general method for cross-classified data without a response variable.

When the response and explanatory variables are both nominal, the method by Goodman and Krusal ([7]) can be further generalized (see Costner [2], and Sarndal [11]). A measure of association of the response variable Y with the explanatory variable X can be defined as

$$(1.1) \quad r(Y|X) = [V(Y) - V(Y|X)]/V(Y),$$

where $V(Y)$ represents a measure of uncertainty in Y without knowledge of X , and $V(Y|X)$ symbolizes the uncertainty in Y when X is known.

The entropy and Gini concentration are most widely used variance mea-

asures for nominal data. Entropy has been widely used in information theory. The Gini concentration has been widely used in statistical analysis of categorical data and economics to measure inequality. Detailed discussions about the entropy and Gini concentrations can be found in Lloyd [8].

The Gini concentration as the measure of variability is defined by

$$(1.2) \quad V_G(X) := \sum_{i=1}^{n_X} p(X = i)(1 - p(X = i)).$$

where n_X represents the number of classes of X .

Combining equations (1.1) and (1.2), the association measure defined by (1.1) reduces to Goodman and Krusal's τ (the GK τ), that is

$$(1.3) \quad \tau^{Y|X} = \frac{\sum_{i=1}^{n_Y} \sum_{j=1}^{n_X} P(Y = i; X = j)^2 / P(X = j) - \sum_{i=1}^{n_Y} P(Y = i)^2}{1 - \sum_{i=1}^{n_Y} P(Y = i)^2},$$

where n_X and n_Y represents the number of classes for the response variable Y and covariate X , respectively. The GK τ is actually equivalent to the conditional Gini index.

We introduce an association vector which measures association at each local response category with an explanatory variable. This vector also provides an expected local accuracy lift rates for proportional prediction. A general class of global association degree is also introduced based on a convex combination of local association measures in the association vector. The coefficients can be chosen according to the objective of the inference. Many measures of association can be derived from the proposed global measure by using different sets coefficients. Moreover, the proposed global measure coincides with the GK τ when the response variable is dichotomous or if the weights are set to be some function of the marginal probabilities of the response variable.

We also propose an association matrix to estimate the generalized confusion matrix before an actual classification. It also provides the distribution of the first and second type like prediction error rates for proportional prediction.

Furthermore, we show that there exists a hierarchy of equivalence relations induced by these measures. This hierarchy provides important insights into the proposed association measures. It is expected to play a crucial role in feature selection and cross classification.

This paper is organized as follows. In Section 2, we introduce the association measure vector and matrix. We also define a general class of association measures by using the proposed association vector. The hierarchy of equivalence relations induced by the association vector, association matrix, and the GK τ is shown in Section 3. We illustrate the properties of the proposed association measures by examining two examples in Section 4. Discussions are provided in Section 5.

2. Association Vector and Matrix. Let X and Y be categorical variables with domains $\text{Dmn}(X) = \{1, 2, \dots, n_X\}$ and $\text{Dmn}(Y) = \{1, 2, \dots, n_Y\}$ respectively. For any $s \in \text{Dmn}(Y)$, we assume that $P(Y = s) > 0$ (thus $P(Y = s) < 1$).

2.1. Association Vector. We introduce an association vector which measures the degrees of each response category associated with an explanatory variable.

DEFINITION 2.1. The *association vector*

$$\Theta^{Y|X} := (\theta^{(Y=1)|X}, \theta^{(Y=2)|X}, \dots, \theta^{(Y=n_Y)|X})$$

is given by

$$(2.1) \quad \theta^{(Y=s)|X} := \frac{E[P(Y=s|X)^2] - P(Y=s)^2}{P(Y=s)(1 - P(Y=s))}, \quad s = 1, 2, \dots, n_Y.$$

When a proportional predictive model based on X is deployed, the components of the vector $\Theta^{Y|X}$ are exactly the error reduction (or accuracy lift) rates for proportional prediction over those using the information of Y only.

PROPOSITION 2.2. *Assume that $P(Y = s) > 0$ for any $s \in \text{Dmn}(Y)$. Then*

- (i) $0 \leq \theta^{(Y=s)|X} \leq 1$;
- (ii) $\theta^{(Y=s)|X} = 0, \quad \forall s \iff Y \text{ and } X \text{ are independent}$;
- (iii) $\theta^{(Y=s)|X} = 1 \iff P(Y = s|X = i) = 1 \text{ or } 0, \text{ for all } i \in \text{Dmn}(X)$.

PROOF. One checks that

$$(2.2) \quad \begin{aligned} \theta^{(Y=s)|X} &= \frac{\sum_{j \in \text{Dmn}(X)} \frac{P(X=j, Y=s)^2}{P(X=j)} - P(Y=s)^2}{P(Y=s)(1 - P(Y=s))} \\ &= \frac{\sum_j (P(X=j, Y=s) - P(X=j)P(Y=s))^2}{P(X=j) P(Y=s)(1 - P(Y=s))} \geq 0, \end{aligned}$$

and

$$(2.3) \quad \begin{aligned} E[P(Y=s|X)^2] &= \sum_j P(Y=s|X=j)^2 P(X=j) \\ &\leq \sum_j P(Y=s|X=j)P(X=j) = P(Y=s). \end{aligned}$$

Thus $\theta^{(Y=s)|X} \leq 1$. The rest follows from (2.2) and (2.3). \square

2.2. *A Class of Global Association Degrees.* One might also be interested in the overall nominal dependence of a response variable on an explanatory variable. Assume that $p(Y = s) > 0$ for all $s = 1, 2, \dots, n_Y (\geq 2)$. We now define a general class of measures for associations.

DEFINITION 2.3. Given a weight vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_{n_Y})$ with $\sum_s \alpha_s = 1$ and $\alpha_s \geq 0$ for all $s = 1, 2, \dots, n_Y$, the global association degree is defined as

$$\tau_{\alpha}^{Y|X} = \sum_{s=1}^{n_Y} \alpha_s \theta^{Y=s|X}.$$

We call $\tau_{\alpha}^{Y|X}$ the α -association degree of Y on X . We call a weight vector α *regular* if $\alpha_s > 0$ for all $s = 1, 2, \dots, n_Y$, in other words, if every single scenario of Y makes contribution to the evaluation of the overall nominal dependence. Sometimes, however, some scenarios of Y may be “merged” with others, e. g. , in the CART decision tree or Value-At-Risk based risk analysis. Therefore the weight vector provides the analyst with a mechanism to place a desired emphasis on certain scenarios given different inferential objectives. In particular, each component of the association vector can be reproduced by placing $\alpha_s = 1$ for a given s .

The following properties of $\tau_{\alpha}^{Y|X}$ follows from Proposition 2.2.

THEOREM 2.4. Assume $\alpha_s > 0$ for all $s \in Dmn(Y)$.

- (i) $0 \leq \tau_{\alpha}^{Y|X} \leq 1$;
- (ii) $\tau_{\alpha}^{Y|X} = 0 \iff Y \text{ and } X \text{ are independent}$;
- (iii) $\tau_{\alpha}^{Y|X} = 1 \iff Y \text{ is completely determined by } X \iff \tau^{Y|X} = 1$.

Given a population the association degree is a function of α and changes accordingly. When employing the proportional prediction model, one can produce the GK τ from the proposed global association degree.

THEOREM 2.5. If the weight vector is assigned as in the following:

$$(2.4) \quad \alpha^P = \frac{1}{V_G(Y)} (P(Y=1) - P(Y=1)^2, \dots, P(Y=n_Y) - P(Y=n_Y)^2);$$

where $V_G(Y) = \sum_s P(Y = s)(1 - P(Y = s))$. we then have

$$(2.5) \quad \tau^{Y|X} = \tau_{\alpha^P}^{Y|X}.$$

PROOF. Observe that

$$\begin{aligned} RHS &= \frac{1}{V_G(Y)} \sum_{i=1}^{n_Y} P(Y = i)(1 - P(Y = i)) \theta^{Y=i|X} \\ &= \frac{1}{V_G(Y)} \sum_{i=1}^{n_Y} (E[P(Y = i)|X]^2 - P(Y = i)^2) \\ &= \frac{1}{V_G(Y)} \left(\sum_{i=1}^{n_Y} \sum_{j=1}^{n_X} P(Y = i; X = j)^2 / P(X = j) - EP(Y) \right). \end{aligned}$$

This completes proof. \square

2.3. Association Matrix. In cross classification with multinomial (instead of binomial) response variable, the association vector essentially shows “correct” classification rates. A complete picture of the cross classification including both correct and detailed misclassifications would be more informative. This motivates the following definition.

DEFINITION 2.6. The association matrix is given by

$$(2.6) \quad \gamma(Y|X) := (\gamma^{st}(Y|X)),$$

where

$$\gamma^{st}(Y|X) := \frac{E[P(Y = s|X) p(Y = t|X)]}{P(Y = s)},$$

where $s, t = 1, 2, \dots, n_Y$.

It can be seen that the association matrix $\gamma(Y|X)$ has the following properties:

- (1) $\gamma(Y|X)$ is a row stochastic matrix;

- (2) $\theta^{(Y=s)|X}$ is the normalization of $\gamma^{ss}(Y|X)$;
- (3) $\gamma^{st}(Y|X) = p(\hat{Y} = t|Y = s, X)$, where \hat{Y} is the predicted value of Y under the proportional prediction and with X as the predictor.

If Y is binary and a proportional prediction is deployed, the matrix $\gamma(Y|X)$ based on the training samples can be explained as the expected confusion-like matrix commonly used in classification; while the conditional probability matrix of $p(\hat{Y}|Y)$ based on actual and predicted responses is a confusion-like matrix. Our proposed association matrix provides the distribution of error rates for proportional prediction.

In addition, the (s, t) -entry, $\gamma^{st}(Y|X)$, of the association matrix $\gamma(Y|X)$, can be regarded as the probability of assigning the true value $Y = s$ to $Y = t$ when the proportional prediction is deployed. Furthermore, when $t \neq s$ and s is fixed, $\gamma^{st}(Y|X)$ is a first-type-like error rate while γ^{ts} is the second-type-like error. If we compute all these error rates based the relationship demonstrated in (Y, \hat{Y}) , it is referred to as the generalized confusion matrix;

More properties of the matrix and relationship with the proposed association vector will be shown in the next section.

3. Hierarchy of Equivalence Relations. In this section, we establish the hierarchy of equivalence relations defined by association matrix, association vector and a global association degree. Recall that for a point set A , a binary relation “ \sim ” on the points of A is referred to as an equivalence relation if it is

- *self-reflective*: $x \sim x$ for all $x \in A$,
- *symmetric*: if $x \sim y$ then $y \sim x$, where $x, y \in A$, and
- *transitive*: if $x \sim y$ and $y \sim z$ then $x \sim z$, where $x, y, z \in A$.

Denote the set of explanatory variables by \mathcal{C} . We now present the five equivalence relations as follows.

DEFINITION 3.1. Let $X_1, X_2 \in \mathcal{C}$ and a response variable Y . With respect to Y , the variables X_1 and X_2 are

3.1.1 *E-1 equivalent*, if $\tau^{X_1|X_2} = \tau^{X_2|X_1} = \tau^{Y|X_1} = 1$;

3.1.2 *E-2 equivalent*, if $\tau^{Y|X_1} = 1 = \tau^{Y|X_2}$;

3.1.3 *E-3 equivalent*, if $\gamma(Y|X_1) = \gamma(Y|X_2)$;

3.1.4 *E-4 equivalent*, if $\Theta^{Y|X_1} = \Theta^{Y|X_2}$;

3.1.5 *E-5 equivalent* with respect to a weight vector α , if $\tau_\alpha^{Y|X_1} = \tau_\alpha^{Y|X_2}$.

THEOREM 3.2. All the above defined binary relations *E-i*, $i = 1, 2, 3, 4, 5$, are indeed equivalence relations on the set \mathcal{C} . Furthermore, if X_1 and X_2 are *E-i* equivalent (with respect to Y), then they are *E-(i+1)* equivalent (with respect to Y), for $i = 1, 2, 3, 4$.

PROOF. Observe that if X_1 and X_2 are E-2 equivalent with respect to Y , then $\gamma(Y|X_i) = I_{n_Y}$, the identity matrix of degree n_Y , $i = 1, 2$.

One checks,

$$(3.1) \quad \gamma^{st}(Y|X) = \sum_{i \in \text{Dmn}(X)} \frac{P(X = i, Y = s)P(X = i, Y = t)}{P(X = i) P(Y = s)}.$$

Thus we have

$$(3.2) \quad \gamma^{ss}(Y|X) = (1 - P(Y = s)) \theta^{Y=s|X} + P(Y = s).$$

Notice that we have

$$\tau_\alpha^{Y|X} = \sum_{s \in \text{Dmn}(Y)} \alpha_s \theta^{Y=s|X}.$$

This completes the proof. \square

From equation (3.2), one can see that the association vector and matrix are functions of each other.

Several remarks related to the equivalence relations are in order.

REMARK 3.3. The E-i equivalence relations depend on the choice of response variable Y . We now show that E-i equivalence is strictly stronger than E-(i+1) equivalence. We are given categorical variables X_1, X_2 and Y in a given data set S .

- a. Considering the following data set S ,

Y	1	0	0	1
X_1	1	2	3	4
X_2	2	3	1	2
probability	2/7	2/7	2/7	1/7

Then we see that X_1, X_2, Y satisfy 3.1.2 but not 3.1.1. Thus 3.1.2 \nRightarrow 3.1.1.

- b. To see in general 3.1.3 \nRightarrow 3.1.2, we notice that if X_1, X_2, Y satisfy 3.1.2, then

$$\gamma^{st}(Y|X_1) = \delta_{st} = \gamma^{st}(Y|X_2)$$

for all $s, t \in \text{Dmn}(Y)$. There exist X_1, X_2, Y satisfying 3.1.3 but not 3.1.2, i. e. ,

$$(\gamma^{st}(Y|X_1)) = (\gamma^{st}(Y|X_2)) \neq (\delta_{st})$$

- c. To see 3.1.4 \nRightarrow 3.1.3 generally, we consider the following data set S .

Y	1	2	2	4	3	4
X_1	1	1	2	2	3	3
X_2	1	3	2	3	1	2
probability	1/6	1/6	1/6	1/6	1/6	1/6

Then

$$\gamma^{qq}(Y|X_i) = \frac{1}{2}, \quad q = 1, 2, 3, 4; \quad i = 1, 2,$$

while

$$\gamma^{12}(Y|X_1) = \frac{1}{2} \neq 0 = \gamma^{12}(X_2).$$

- d. To see 3.1.5 \nRightarrow 3.1.4 generally, we consider the following data set S .

Y	1	1	2	3	1	2	3	2
X_1	1	1	2	3	4	1	1	4
X_2	2	1	1	1	4	1	3	4
probability	1/10	2/10	1/10	1/10	1/10	2/10	1/10	1/10

Then

$$\begin{aligned} \tau^{Y|X_1} &= \tau^{Y|X_2} = \frac{9}{25}, \\ (\theta^{Y=1|X_1}, \theta^{Y=2|X_1}, \theta^{Y=3|X_1}) &= \left(\frac{1}{6}, \frac{17}{72}, \frac{23}{48}\right), \\ (\theta^{Y=1|X_2}, \theta^{Y=2|X_2}, \theta^{Y=3|X_2}) &= \left(\frac{17}{72}, \frac{1}{6}, \frac{23}{48}\right). \end{aligned}$$

- e. If we replace E-2 equivalence with E-2', where E-2' is defined as “ X_1 and X_2 are E-2' equivalent if $\tau^{X_1|X_2} = 1 = \tau^{X_2|X_1}$ ”, then the stronger-to-weaker chain that

$$\text{E-1} \implies \text{E-2}' \implies \text{E-3} \implies \text{E-4} \implies \text{E-5} \text{ (but not vice versa)}$$

still holds.

To see that E-2' implies E-3, notice that since $\tau^{X_1|X_2} = 1 = \tau^{X_2|X_1}$, we have that $|\text{Dmn}(X_1)| = |\text{Dmn}(X_2)|$ and for any event $X_1 = i$ there is a unique $X_2 = j$ such that $P(X_2 = j|X_1 = i) = 1$ and vice versa. Assume that $\text{Dmn}(X_1) = \{i_1, \dots, i_k\}$. Then $\text{Dmn}(X_2) = \{j_1, \dots, j_k\}$ and we may and shall assume that

$$P(X_2 = j_q|X_1 = i_q) = 1 = P(X_1 = i_q|X_2 = j_q), \quad q = 1, \dots, k.$$

Thus

$$\begin{aligned} \gamma^{st}(Y|X_1) &= \sum_{q=1}^k \frac{P(X_1 = i_q, Y = s) P(X_1 = i_q, Y = t)}{P(X_1 = i_q) P(Y = s)} \\ &= \sum_{q=1}^k \frac{P(X_2 = j_q, Y = s) P(X_2 = j_q, Y = t)}{P(X_2 = j_q) P(Y = s)} \\ &= \gamma^{st}(Y|X_2). \end{aligned}$$

Thus X_1 is E-3 equivalent to X_2 . It is easy to see in general E-3 equivalence does not imply E-2' equivalence.

f. E-1 equivalence condition can be replaced with “if $\tau_\alpha^{X_1|X_2} = \tau_\alpha^{X_2|X_1} = \tau_\alpha^{Y|X_1} = 1$ for a regular weight vector α ”.

Actually, $\tau_\alpha^{X_1|X_2} = \tau_\alpha^{X_2|X_1} = \tau_\alpha^{Y|X_1} = 1$ if and only if $\tau^{X_1|X_2} = \tau^{X_2|X_1} = \tau^{Y|X_1} = 1$.

Similarly, E-2 equivalence condition can be replaced with “if $\tau_\alpha^{Y|X_1} = 1 = \tau_\alpha^{Y|X_2}$ for a regular weight vector α ”.

The equivalence relations and the hierarchy are expected to lay a foundation for feature selection and prediction for categorical data.

In clinical trials, direct/search marketing or risk management, one also often faces the case of binary response variable. We shall see from the following theorem that the five relations actually degenerate to three.

THEOREM 3.4. *If the response variable is dichotomous, say, $Dmn(Y) = \{0, 1\}$, then for any weight vector α*

- (1) $\theta^{Y=s|X} = \tau_\alpha^{Y|X}, s = 0, 1;$
- (2) *with respect to Y , the E - i equivalence relations are the same for $i = 3, 4, 5$.*

PROOF. (1). It is a routine check.

(2). One checks

$$\begin{aligned}\gamma^{11}(Y|X) &= P(Y = 1) + \frac{V_G(Y) \tau_\alpha^{Y|X}}{2P(Y = 1)}, \\ \gamma^{22}(Y|X) &= P(Y = 2) + \frac{V_G(Y) \tau_\alpha^{Y|X}}{2P(Y = 2)};\end{aligned}$$

by (1), we have $\gamma^{11}(Y|X) = \gamma^{22}(Y|X)$. Notice that $\gamma(y|x)$ is a two-by-two row stochastic matrix. So the matrix is uniquely determined by $\tau_\alpha^{Y|x}$. \square

This theorem implies that τ_α (in particular, the GK τ) is enough in the case of binary response variable.

The following theorem tells us about what the joint distribution (X_1, X_2) looks like if X_1 and X_2 are E -2 equivalent with respect to Y .

THEOREM 3.5. *X_1 and X_2 are E -2 equivalent w. r. t. Y if and only if Y is completely determined by X_1 or X_2 ; and in this case, there exist hard partitions $Par(X_i) := \{X_i^s | s \in Dmn(Y)\}$ of $Dmn(X_i)$, $i = 1, 2$, where each X_i^s consists of some scenarios of X_i , such that*

$$\{(a_1, a_2) | a_1 \in Dmn(X_1^s), a_2 \in X_2^t\} = \emptyset \text{ whenever } s \neq t;$$

while

$$\{(a_1, a_2) | a_1 \in Dmn(X_1^s), a_2 \in Dmn(X_2^s)\} \neq \emptyset$$

for all $s \in Dmn(Y)$.

PROOF. Without loss of generality, we may and shall assume $\text{Dmn}(Y) = \{1, 2, \dots, n_Y\}$. Since $\tau^{Y|X_i} = 1$, the pairs of (X_i, Y) in S defines naturally a deterministic surjective function $f_i : \text{Dmn}(X_i) \rightarrow \text{Dmn}(Y)$ so that $P(Y = f(a_i)|X_i = a_i) = 1$. Thus $X_i^p := f_i^{-1}(p)$, $p = 1, \dots, n_Y$, defines a partition of $\text{Dmn}(X_i)$. We want to show that

$$S(p, p) := \{(a_1, a_2) | a_i \in f_i^{-1}(p), \quad i = 1, 2\} \neq \emptyset,$$

while when $1 \leq p \neq q \leq n_Y$,

$$S(p, q) := \{(a_1, a_2) | a_1 \in f_1^{-1}(p), \quad a_2 \in f_2^{-1}(q)\} = \emptyset.$$

In fact, let $a_1 \in f_1^{-1}(p)$. Then there exists $a_2 \in \text{Dmn}(X_2)$ such that $(a_1, a_2, p) \in X_1 \times X_2 \times Y$. Thus $a_2 \in f_2^{-1}(p)$. Hence $S(p, p) \neq \emptyset$. By the arbitrariness of a_1 in $f_1^{-1}(p)$, we see that $S(p, q) = \emptyset$, when $p \neq q$.

The above argument also shows that

$$X_1 \times X_2 = \cup_{p \in \text{Dmn}(Y)} S(p, p).$$

This completes proof. \square

4. Examples and Data Analysis. To illustrate, we present two examples including analysis results from a credit risk management data set.

EXAMPLE 4.1. We first consider a data set with 23,370 observations. The response variable has 6 scenarios with the following probability distribution:

$$p(Y) = (.1048, .3083, .3062, .1563, .1092, .0142).$$

(i) To demonstrate, we generate another categorical variables which is independent of the response variable. The generated explanatory variable has 6 categories. Since the response and explanatory variables are independent,

$X \setminus Y$	1	2	3	4	5	6	(X, \cdot)
1	788	183	0	0	0	0	971
2	1089	4358	3006	800	160	0	9412
3	320	1665	2544	1558	1101	55	7242
4	131	559	949	746	660	97	3142
5	92	363	583	493	522	141	2194
6	29	78	75	78	109	40	409
(\cdot, Y)	2449	7205	7156	3675	2552	333	23370

TABLE 1
Joint frequency table.

every component of the association vector should also be zero in theory. The global α -association degree should be zero for any given weight vector α . The estimated association vector is calculated to be

$$\Theta = (2 \times 10^{-3}, 2 \times 10^{-3}, 3 \times 10^{-3}, 10^{-3}, 5 \times 10^{-3}, 3 \times 10^{-3}).$$

(ii) Next we select an actual explanatory variable contained in the same data set. The joint frequency table is given in Table 4.1. The GK τ and the association vector are shown as follows:

$$\tau = .0763; \quad \Theta = (.2437, .0778, .0236, .0413, .0806, .0355).$$

This shows that the selected actual explanatory variable has some demonstrated mild association with the response variable.

(iii) In order to demonstrate the properties of the association matrix, we randomly select 80% of the observations and put them into the training set. The rest is set aside for test set.

The association vectors restricted to the training and test sets are calculated:

$$\begin{aligned} \Theta_{train} &= (.2348, .1369, .0457, .0374, .0500, .0158); \\ \Theta_{test} &= (.2502, .1410, .0532, .0336, .0608, .0121). \end{aligned}$$

By using the same proportional prediction principle described in Goodman and Kruskal [7], one can make a guess on the category of the response variable for each observation based on the conditional distributions, and then obtain the generalized confusion matrix.

The association matrix based on training (left) and the generalized confusion matrix on validation (right) are given as:

$$\begin{pmatrix} .26 & .48 & .15 & .06 & .04 & .01 \\ .05 & .49 & .28 & .11 & .07 & .01 \\ .02 & .37 & .34 & .15 & .11 & .02 \\ .02 & .32 & .35 & .17 & .12 & .02 \\ .02 & .30 & .34 & .17 & .14 & .03 \\ .03 & .28 & .33 & .18 & .15 & .03 \end{pmatrix} VS \begin{pmatrix} .23 & .50 & .18 & .04 & .04 & .01 \\ .05 & .48 & .27 & .11 & .07 & .01 \\ .01 & .35 & .35 & .16 & .11 & .02 \\ .02 & .33 & .33 & .17 & .11 & .03 \\ .01 & .28 & .37 & .16 & .13 & .04 \\ .06 & .24 & .33 & .18 & .18 & .01 \end{pmatrix}$$

It can be seen that the generalized confusion matrix by using the test set is very close to the association matrix by using the training set. Since both matrices are row-stochastic matrices, we test the hypothesis that the two matrices form two identical distributions. The hypothesis was not rejected since the p-value is very close to 1.

EXAMPLE 4.2. We now consider a real loan application data discussed used in [12] or [10]. This data set has several variables and 650 records. For simplicity, we are only concerned about these five (categorical or discretized) variables: *On-Time*, *Age*, *Income*, *Credit* and *Risk*, where each variable was categorized as *On-Time*=(No (0), Yes (1)); *Age*=(young, med, sen); *Income*=(low, mid, hi); *Risk*=(low, med, hi); *Credit*=(red, yellow, green). We consider three situations in which *On-Time*, *Risk*, and *Credit* are used as the response variable respectively.

1. For response $Y = \text{On-Time}$, we observe that $p(Y) = (0.1, 0.9)$. Since Y is binary, by Theorem 3.4, $\tau_\alpha^{Y|X} = \tau^{Y|X} = \theta^{Y=i|X}, i = 0, 1$.

X	Credit	Risk	Age	Income
$\tau^{Y X}$.0577	.0486	.0402	.0134

2. For response $Y = \text{Risk}$: $p(Y) = (.4877, .0400, .4723)$, we obtain the following results:

X	$\tau^{Y X}$	$\Theta^{Y X}$	$\gamma(Y X)$	(X,Y) freq.
On-Time	.0432	(.0451,.0002,.0479)	$\begin{pmatrix} .5108 & .0407 & .4485 \\ .4959 & .0402 & .4639 \\ .4631 & .0393 & .4976 \end{pmatrix}$	$\begin{pmatrix} 11 & 2 & 52 \\ 306 & 24 & 255 \end{pmatrix}$
Age	.5137	(.5451,.0018,.5611)	$\begin{pmatrix} .7669 & .0437 & .1894 \\ .5324 & .0417 & .4258 \\ .1956 & .0361 & .7684 \end{pmatrix}$	$\begin{pmatrix} 13 & 9 & 246 \\ 291 & 17 & 61 \\ 13 & 0 & 0 \end{pmatrix}$
Income	.0272	(.0368,.0207,.0185)	$\begin{pmatrix} .5065 & .0345 & .459 \\ .4206 & .0599 & .5195 \\ .4739 & .044 & .4821 \end{pmatrix}$	$\begin{pmatrix} 19 & 8 & 45 \\ 211 & 17 & 209 \\ 87 & 1 & 53 \end{pmatrix}$
Credit	.0009	(.0006,.0008,.0012)	$\begin{pmatrix} .488 & .0401 & .4719 \\ .4892 & .0408 & .4700 \\ .4872 & .0398 & .4729 \end{pmatrix}$	$\begin{pmatrix} 35 & 2 & 40 \\ 98 & 9 & 93 \\ 184 & 15 & 174 \end{pmatrix}$

3. For response $Y = \text{Credit}$: $p(\cdot, Y) = (.1185, .3077, .5738)$, we obtain the following results:

X	$\tau^{Y X}$	$\Theta^{Y X}$	$\gamma(Y X)$	(X,Y) freq.
On-Time	.0319	(.0322, .0123, .0488)	$\begin{pmatrix} .1468 & .3328 & .5204 \\ .1281 & .3162 & .5556 \\ .1074 & .2979 & .5946 \end{pmatrix}$	$\begin{pmatrix} 19 & 30 & 16 \\ 58 & 170 & 357 \end{pmatrix}$
Age	.0035	(.0099, .0028, .0014)	$\begin{pmatrix} .1272 & .3023 & .5705 \\ .1164 & .3096 & .5740 \\ .1178 & .3078 & .5744 \end{pmatrix}$	$\begin{pmatrix} 40 & 80 & 148 \\ 34 & 118 & 217 \\ 3 & 2 & 8 \end{pmatrix}$
Income	.001	(.0007, .0006, .0016)	$\begin{pmatrix} .1191 & .3085 & .5724 \\ .1188 & .3081 & .5731 \\ .1182 & .3073 & .5745 \end{pmatrix}$	$\begin{pmatrix} 7 & 20 & 45 \\ 54 & 137 & 246 \\ 16 & 43 & 82 \end{pmatrix}$
Risk	.0005	(.0016,.0003,.0002)	$\begin{pmatrix} .1199 & .3069 & .5733 \\ .1181 & .3079 & .5739 \\ .1183 & .3077 & .5739 \end{pmatrix}$	$\begin{pmatrix} 35 & 98 & 184 \\ 2 & 9 & 15 \\ 40 & 93 & 174 \end{pmatrix}$

The variable *Risk* was generated by a seemingly subjective discretization on the ratios of debt over asset and set to reflect the degree of risk of the loan or borrower. Calculations have shown that *Risk* and *Credit* are almost independent of each other. Moreover, the variable *On-Time* is quite lowly associated with each of the two variables.

We think that there are two main reasons: (1) either the credit scoring or the risk assigned is in poor quality; or even both are in poor quality. In this case, the continuous variable should be properly discretized. (2) The existing categorized Risk is a subjective or conventional classification of the debt-over-asset ratios. We can see that this classification is almost meaningless for the loan risk management.

5. Discussions and Future Work. We introduce an association vector, a class of global association degrees based on the association vector and weight vectors, and an association matrix. We also study the equivalence relations induced by the association measures introduced.

The association vector measures both local-on-global and global-on-global nominal cross-classification dependence. One may directly see the expected proportional prediction's lifts for each value of the response variable. The association vector is essentially equivalent with the diagonal the association matrix. Various global associations can be derived from the association vector with various weights. Our subsequent work on feature selection algorithms will be based on the class of global association measures proposed in this article.

The association matrix based on proportional classification gives rise to the accuracy rate and error rate distribution. It is an extension of the confusion matrix widely used in classification. We expect more research on relevant statistical inference and variations of the association matrix in the future.

The hierarchy helps deepen understanding of local and global statistical association among variables and the structure of a multivariate distribution. The equivalence relations are expected to play a crucial role in analysis of

high dimensional categorical data when the populations are not known and sample sizes are small.

References.

- [1] Agresti, A. (2002). *Categorical Data Analysis*, John Wiley, New York.
- [2] Costener, H.L. (1965). Criteria for measure of association, *American Sociological Review*, **30**, 341-353.
- [3] Fienberg, S. (2007). *The Analysis of Cross-Classified Categorical Data*, Springer, New York.
- [4] Gini, C.W. (1971) Variability and Mutability, contribution to the study of statistical distributions and relations, *Studi Economico-Giuridici della R. Universita de Cagliari* (1912). Reviewed in: Light, R.J., Margolin, B.H.: *An Analysis of Variance for Categorical Data*. *Journal of the American Statistical Association*, **66**, 534-544.
- [5] Goodman, L.A. (1996). A Single General Method for the Analysis of Cross-Classified Data: Reconciliation and Synthesis of Some Methods of Pearson, Yule and Fisher, and Also Some Methods of Correpondence Analsysis and Association Analysis, *The Journal of the American Statistical Association*, **91**, 408-428.
- [6] Goodman, L.A. (2000). The analysis of cross-classified data: notes on a century of progress in contingency table analysis, and some comments on its prehistory and its future, *Statistics for the 21st Century*, editors: Rao, C.R. and Szekely, G. J., 189-231, Marcel Dekker.
- [7] Goodman, L.A. and Kruskal, W. H. (1954). Measures of Associations for Cross classification, *The Journal of the American Statistical Association*, **49**, 732-764.
- [8] Lloyd, C. J. (1999). *Statistical analysis of categorical data*, John Wiley & Sons.
- [9] Micheli-Tzanakou, E. (1999). *Supervised and unsupervised pattern recognition: feature extraction and computational*, CRC Press.
- [10] Olson, D. and Shi, Y. (2007). *Introduction to business data mining*, McGraw-Hill.
- [11] Sarndal, C.E. (1974). A comparative study of association measures, *Psychometrika*, **39**, 165-187.
- [12] Seppanen, M. S., Kumar S. and Chandra, C. (2004) *Process Analysis and Improvement: Tools and Techniques*, McGraw-Hill Higher Education.

WENXUE HUANG

DEPARTMENT OF MATHEMATICS

SHANTOU UNIVERSITY

P.R. CHINA

E-MAIL: wxhuang@stu.edu.cn

YONG SHI

RESEARCH CENTER FOR FICTITIOUS ECONOMICS AND DATA SCIENCE

CHINESE ACADEMY OF SCIENCE

P.R. CHINA

E-MAIL: yshi@gucas.ac.cn

XIAOGANG(STEVEN) WANG

DEPARTMENT OF MATHEMATICS AND STATISTICS

YORK UNIVERSITY

CANADA

E-MAIL: stevenw@mathstat.yorku.ca